Radiology

Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI

*Daniel Truhn, MS, MD** • *Simone Schrading, MD** • *Christoph Haarburger, MS* • *Hannah Schneider, BS* • *Dorit Merhof, PhD* • *Christiane Kuhl, MD*

From the Departments of Diagnostic and Interventional Radiology (D.T., S.S., H.S., C.K.) and Institute of Imaging and Computer Vision (C.H., D.M.), RWTH Aachen University, Aachen, Pauwelsstr 30, 52074 Aachen, Germany. Received June 9, 2018; revision requested July 26; revision received September 25; accepted September 26. **Address correspondence to** C.K. (e-mail: *ckuhl@ukaachen.de*).

D.T. supported by the START research grant of the medical faculty of the University of Aachen.

*D.T. and S.S. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

Radiology 2019; 290:290–297 • https://doi.org/10.1148/radiol.2018181352 • Content codes: MR IN BR

Purpose: To compare the diagnostic performance of radiomic analysis (RA) and a convolutional neural network (CNN) to radiologists for classification of contrast agent–enhancing lesions as benign or malignant at multiparametric breast MRI.

Materials and Methods: Between August 2011 and August 2015, 447 patients with 1294 enhancing lesions (787 malignant, 507 benign; median size, 15 mm ± 20) were evaluated. Lesions were manually segmented by one breast radiologist. RA was performed by using L1 regularization and principal component analysis. CNN used a deep residual neural network with 34 layers. All algorithms were also retrained on half the number of lesions (n = 647). Machine interpretations were compared with prospective interpretations by three breast radiologists. Standard of reference was histologic analysis or follow-up. Areas under the receiver operating curve (AUCs) were used to compare diagnostic performance.

Results: CNN trained on the full cohort was superior to training on the half-size cohort (AUC, 0.88 vs 0.83, respectively; P = .01), but there was no difference for RA and L1 regularization (AUC, 0.81 vs 0.80, respectively; P = .76) or RA and principal component analysis (AUC, 0.78 vs 0.78, respectively; P = .93). By using the full cohort, CNN performance (AUC, 0.88; 95% confidence interval: 0.86, 0.89) was better than RA and L1 regularization (AUC, 0.81; 95% confidence interval: 0.79, 0.83; $P < .001$) and RA and principal component analysis (AUC, 0.78; 95% confidence interval: 0.76, 0.80; $P < .001$). However, CNN was inferior to breast radiologist interpretation (AUC, 0.98; 95% confidence interval: 0.96, 0.99; $P < .001$).

Conclusion: A convolutional neural network was superior to radiomic analysis for classification of enhancing lesions as benign or malignant at multiparametric breast MRI. Both approaches were inferior to radiologists' performance; however, more training data will further improve performance of convolutional neural network, but not that of radiomics algorithms.

Online supplemental material is available for this article.

MRI is a powerful tool for diagnosis and screening of breast cancer (1). However, widespread use of breast MRI has been restricted because of the limited availability of sites that offer this method. One major reason for limited availability is the lack of radiologists who can offer substantial expertise in interpreting breast MR images.

Sophisticated machine learning approaches show promise in complementing human diagnosis (2). Broadly speaking, machine learning can be divided into two major classes: one is radiomic analysis (RA), where hand-made image features are extracted; and the other is the concept of convolutional neural networks (CNN), in which the computer learns to recognize image features on its own, usually on the basis of a set of labeled training examples. Both approaches have been pursued with considerable success for image interpretation, although in different areas: In the field of diagnostic radiology, RA has been successfully used to further classify tumor types (3,4). However, CNNs require a larger pool of training images before they achieve a clinically useful performance.

Within radiology, breast imaging, specifically mammographic screening, lends itself to be used with CNNs because similarly large data sets are available (5,6). With such large mammographic data sets, and with the advent of increased computing power, deep learning may have the potential to outperform regular computer-assisted diagnosis systems for mammographic interpretation (5).

Studies are limited regarding the use of RA or CNNs for diagnostic classification of contrast agent–enhancing breast lesions (ie, for differential diagnosis of benign vs malignant lesions). Bickelhaupt et al (7) used machine learning for further characterization of lesions suspicious for cancer that were found on digital mammographic images and used unenhanced and diffusion-weighted MRI for this purpose. However, the use of RA or CNNs for classification of enhancing lesions observed at regular, clinical, dynamic contrast agent–enhanced, or multiparametric breast MRI is not established.

Because breast MRI is performed less than mammographic screening, available breast MRI data sets are smaller

Abbreviations

AUC = area under the curve, BI-RADS = Breast Imaging Reporting and Data System, CNN = convolutional neural network, RA = radiomic analysis

Summary

Convolutional neural networks were superior to radiomic analysis for classification of enhancing lesions. Although both approaches were inferior to radiologists' performance, convolutional neural networks have the potential for further improvement with more training data whereas radiomic analysis does not.

Implications for Patient Care

- For machine interpretation of multiparametric MRI examinations in the breast, convolutional neural networks seem to be more suitable than radiomic analysis.
- Performance of radiologists was superior compared with both radiomic analysis and convolutional neural networks (area under the curve, 0.98 versus 0.81 and 0.88, respectively.
- With more training data, convolutional neural networks have the potential to further improve their performance and close the gap to human performance.

Table 1: Patient Demographics

Parameter	Result
Mean age (y)*	64 (26–82)
Median	66
Sex	
Women	447 (100)
Men	0 (0)
Menopausal status	
Before menopause	110 (24.5)
After menopause	339 (75.5)
Risk status	
Personal history of breast cancer	60 (13.4)
Family history of breast cancer	48 (10.7)
Personal history of breast biopsy with atypias	18 (4.0)
Current diagnosis of breast cancer	243 (54.4)
Average risk	78 (17.4)

Note.—Unless otherwise indicated, data are number of patients and data in parentheses are percentages.

* Data in parentheses are range.

than the current data sets used for CNN analyses of medical (and nonmedical) images. Considering this, and the more complex nature of dynamic contrast-enhanced MRI compared with digital mammography (for example), we were interested to find out how CNNs performed compared with RA by using MRI data-set volumes that are attainable in a clinical setting.

Accordingly, the aim of our study was to determine the performance of three different machine learning algorithms (two variants of radiomic analyses, L1 regularization and principal component analysis, and convolutional neuronal networks) compared with the performance of radiologists for classification of enhancing lesions at multiparametric dynamic contrast-enhanced MRI.

Materials and Methods

Local institutional review board approval was obtained. Patients provided written informed consent to have their imaging data analyzed.

Our study evaluated breast MRI examinations performed between August 2011 and August 2015 and prospectively interpreted by radiologists in an academic breast center.

To generate the analysis cohort, first we selected from our picture archiving and communication system patients who had undergone breast MRI in our department between August 2011 and August 2015. This search yielded 5687 breast MRI examinations. From this cohort, we randomly retrieved 1000 patients. We then excluded breast MRI studies that did not fulfill the following criteria: breast MRI studies that *(a)* did not exhibit an enhancing lesion and/or that *(b)* had no validation of the final diagnosis, and/or that *(c)* could not be clearly allocated to either of the binary classification categories (ie, benign or malignant) (Fig 1). Validation was achieved either by histopathologic analysis (all lesions categorized at MRI as Breast Imaging Reporting and Data System [BI-RADS] category 4, 5, or 6) or by an uneventful MRI follow-up of at least 24 months (for all lesions categorized at MRI as BI-RADS category 2 or 3). To provide an unambiguous ground truth of diagnoses (either malignant or benign), we excluded patients with borderline (ie, high-risk) lesions.

To avoid repeated observations in the same patients that would confound the results because of interlesion correlation, in patients with several enhancing lesions of the same type (eg, multicentric cancer or multiple fibroadenoma), we included only one type of enhancing lesion per breast (eg, only one of the invasive cancers).

Splitting into training, validation, and analysis data for all algorithms was performed in a patient-wise fashion with a 10-fold cross validation in the outer loop and a fivefold cross validation in the inner loop, resulting in a 72%/18%/10% split into independent training, validation, and analysis sets, respectively. Each of the 10 folds in the outer loop results in a score denoting the probability of malignancy allocated by the algorithm for 10% of the lesions. Because the analysis sets between the folds are disjoint and their union covers the whole set, we arrived at scores for all lesions. A more detailed description of the splitting process is provided in Appendix E1 (online).

Multiparametric contrast-enhanced bilateral breast MRI had been performed according to a previously published standardized protocol (8). In short, the protocol consists of an axial bilateral T2-weighted fast spin-echo and an axial bilateral dynamic series consisting of five dynamic phases (one phase before administration of contrast agent and four postcontrast phases) acquired without fat suppression. Image subtraction was performed at all postcontrast phases. Breast MRI studies were interpreted prospectively in about equal proportions by three different readers (D.T., S.S., and C.K., with between 7 and 25 years of experience in interpreting breast MRI studies). In addition to the usual overall BI-RADS categorization, MRI reports list BI-RADS categories also on a per-lesion basis to facilitate communication about management of specific lesions across different breast imaging modalities.

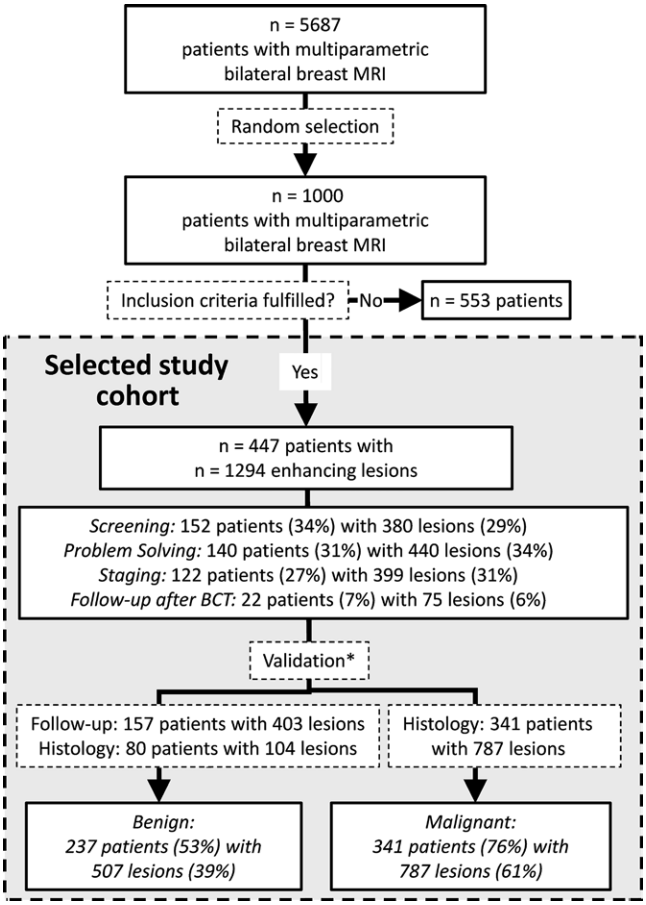

Figure 1: Flowchart of the final analysis cohort. Exclusion criteria were breast MRI studies that did not exhibit an enhancing lesion and/or that had no validation of the final diagnosis, and/or that could not be clearly allocated to either of the binary classification categories (ie, benign or malignant). BCT = breast-conserving treatment. *Validation was obtained per lesion; because patients may have more than one lesion, the total number of patients exceeds the total number of patients included in the analysis cohort.

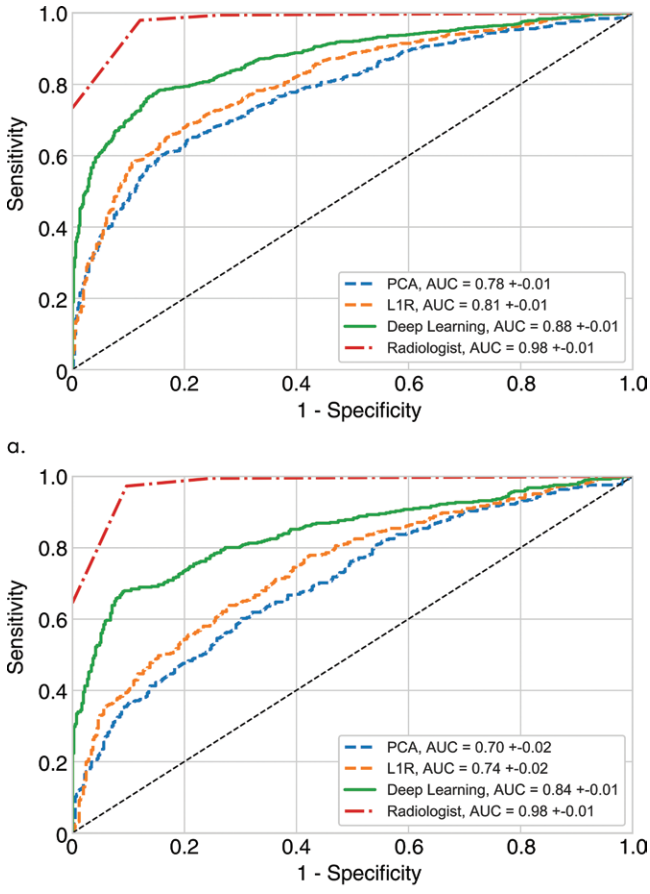

Figure 2: Receiver operating characteristic curve analysis for radiomic analysis (RA) with L1 regularization (L1R) for RA with principal component analysis (PCA) for the convolutional neural network (CNN) and radiologist for **(a)** all lesions in data set and **(b)** the subset of lesions that are smaller than 2 cm. In all cases, the CNN performed better ($P < .001$) than did both conventional machine learning RA algorithms (L1, L1R; PCA). Radiologists' readings were superior to all three algorithms ($P < .001$).

All computations were performed on a desktop computer equipped with Intel Core i7–7700 K processor (Intel, Santa Clara, Calif) and Nvidia GTX 1080 Ti GPU (Nvidia, Santa Clara, Calif). When not otherwise specified, code implementations were in-house developments based on python 3.6.5 (*https://www.python.org*) and the software modules numpy, scipy, and sklearn (9).

Lesions were manually segmented by a single breast radiologist (S.S., with 15 years of experience interpreting breast MRI studies). To segment, the radiologist first reviewed all images to identify the ones that were best suited to view lesion boundaries. Segmentation was performed either on the subtracted images (in case there was no motion) or on the nonsubtracted source images in case subtraction errors because of motion were present. Lesions were segmented on a section-by-section basis until the full lesion volume was captured and a three-dimensional lesion volume was acquired. Next, the region of interest was propagated to all remaining sequences on which the lesion had not directly been segmented. To ensure comparability of image signal intensities across patients, a bias field correction with N4ITK (10) was executed on all images and image intensities were rescaled to a fixed range of 0–511.

For each lesion, statistical, shape, and texture features were extracted by using the Pyradiomics (11) toolbox. The distribution of image intensities within a lesion were quantified by 19 statistical features. Extracted texture features were calculated on Gray-level co-occurrence matrix (27 features), Gray-level run-length matrix (16 features), and Gray-level size-zone matrix (16 features). Both statistical and texture features were extracted for the T2-weighted image, the subtraction image of the first (precontrast) dynamic acquisition, and the four postcontrast dynamic acquisitions, respectively, resulting in a total of 133 statistical and 413 texture features. In total, 16 shape features assessed spatial properties of lesions and were extracted on the basis of the segmentation mask. A detailed definition of all image features can be found online (*http://pyradiomics.readthedocs.io/en/latest/features.html*).

To select a suitable subset of features that was both limited in size and uncorrelated, the following two distinct feature selection strategies were evaluated: *(a)* L1 regularization: features were selected implicitly by L1 regularization of the linear classifier. L1 regularization enforces linear model's coefficient

to have a sparse solution, leading to a small subset of chosen features in the model; and *(b)* principal component analysis in which a subset of 100 features was selected on the basis of their power in differentiating between malignant and benign lesions in the training set (ie, by exhibiting the lowest *P* values), and subsequently, principal component analysis was performed on those features and the resulting first 10 principal components were used as inputs. Details about the feature selection strategies and the use of hardware and software are provided in Appendix E1 (online) and Tables E1 and E2 (online).

We used a network architecture that was previously described (12). In short, a deep residual neural network (ResNet18) (13) was pretrained on a dataset of 14 000 000 color photographs of everyday objects (14) to sensitize the deep layers to potentially relevant structural information such as edges and lines. Details of the network architecture are provided in Appendix E1 (online). Data augmentation was performed by using random rotations and flipping. We used stochastic gradient descent with

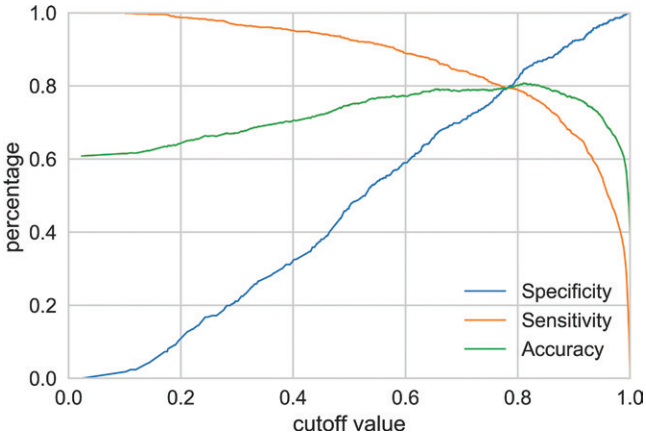

Figure 3: Graph of the sensitivity and specificity depending on cut-off value chosen for classification of convolutional neural network outputs. The convolutional neural network gives a numeric value between 0 and 1 that can be regarded as a pseudoprobability for malignancy. Depending on the chosen cutoff that should indicate a positive or negative test result, values for sensitivity, specificity, and accuracy vary. The cutoff was chosen to minimize $(1 - \text{sensitivity})^2 \times (1 - \text{specificity})^2$, which resulted in a cutoff value of 0.81. Sensitivities and specificities were calculated on the basis of the 507 benign and 787 malignant lesions in the dataset.

Table 2: Description of Lesions Included in the Data Set

Lesion type	No. of Lesions
Benign lesions	507 (39.2)
Adenosis	156 (30.8)
Fibroadenoma	153 (30.2)
Lymph node	63 (12.4)
Papilloma	20 (3.9)
Other*	115 (22.7)
Malignant lesions	787 (60.8)
Ductal carcinoma in situ	253 (32.1)
Invasive cancer	534 (67.9)
Type of enhancement	
Enhancing masses	704 (55.4)
Benign	226 (32.1)
Malignant	478 (67.9)
Non mass enhancement	590 (45.6)
Benign	226 (38.3)
Malignant	364 (61.7)
Mean size of lesions (mm)†	
Overall	22.4 ± 20.3 (4.0–131.2)
Malignant	27.0 ± 22.7 (4.2–131.2)
Invasive	20.2 ± 14.3 (4.2–99.4)
Ductal carcinoma in situ	40.9 ± 29.0 (4.3–131.2)
Benign	15.3 ± 13.2 (4.0–85.9)

Note.—There were a total of 1294 lesions. Unless otherwise indicated, data in parentheses are percentage.

* The "other" category included enhancement around fat necrosis, fresh scar tissue, pseudoangiomatous stromal hyperplasia, and other benign-appearing enhancement because of focal or regional background enhancement.

† Data are ± standard deviation; data in parentheses are range.

Table 3: Diagnostic Performances of Radiomic Analysis by Using L1 Regularization, Radiomic Analysis with Principal Component Analysis, and Convolutional Neural Networks

	Radiomic Analysis			
Parameter	L1 Regularization	Principal Component Analysis	Convolutional Neural Network	Breast Radiologist
Sensitivity (%)	71.5 (563/787) [69.0, 74.0]	67.9 (534/787) [65.4, 70.4]	78.3 (616/787) [76.1, 80.5]	99.7 (785/787) [99.4, 100]
Specificity (%)	76.1 (386/507) [73.8, 78.4]	75.1 (381/507) [72.7, 77.5]	84.6 (429/507) [82.6, 86.6]	86.4 (438/507) [84.5, 88.3]
Area under the curve with full-size data set	0.81 [0.79, 0.84]	0.78 [0.75, 0.80]	0.88 [0.86, 0.89]	0.98 [0.97, 0.99]
Area under the curve with half-size data set	0.80 [0.77, 0.84]	0.78 [0.74, 0.82]	0.83 [0.79, 0.86]	0.96 [0.95, 0.98]

Note.—Data in parentheses are numerator/denominator; data in brackets are 95% confidence intervals. Table includes half-size and full-size training cohort. For calculation of sensitivities and specificities of the machine learning algorithms, it was necessary to choose a cutoff value (see Fig 3). The value was chosen to minimize the distance to the upper left corner in the receiver operating characteristic curve value of 0.81.

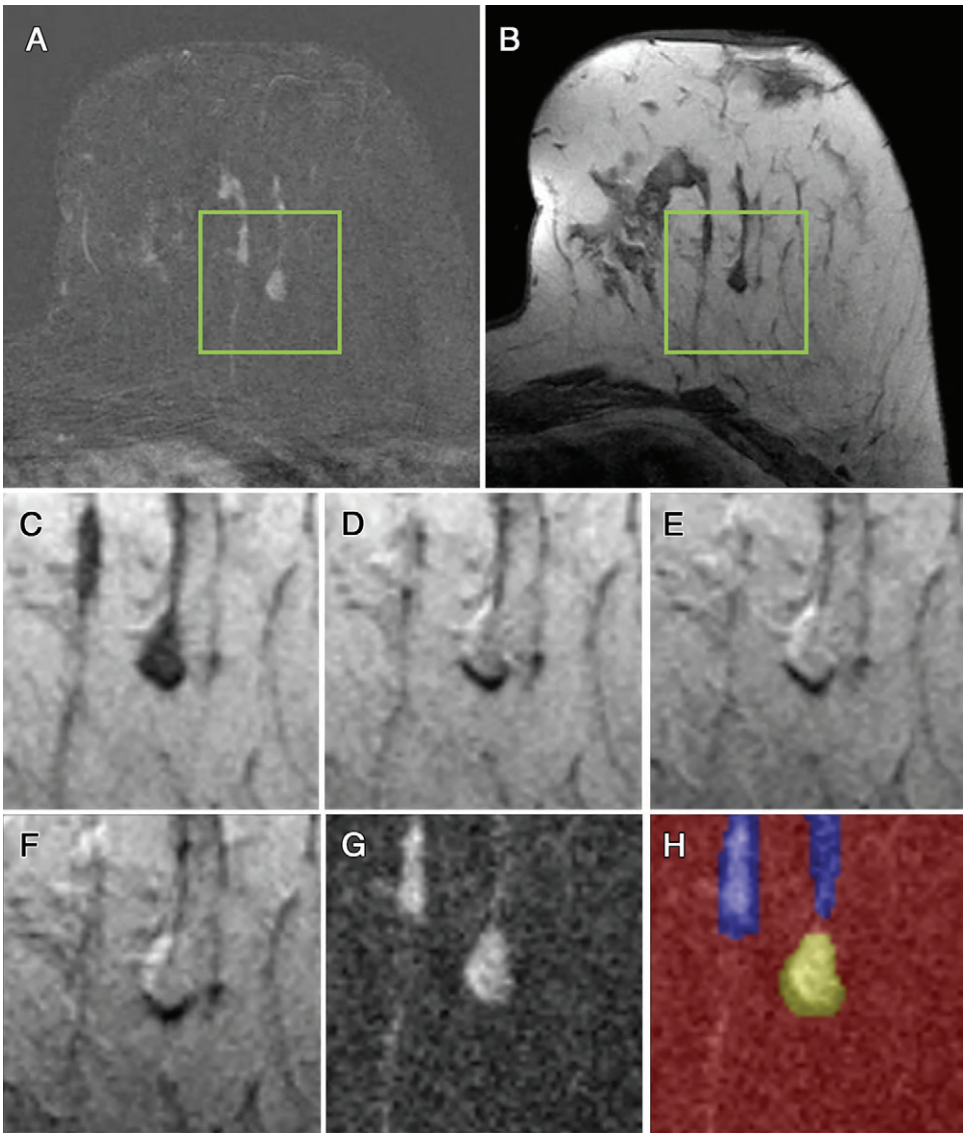

Figure 4: A breast lesion in a 53-year-old female patient in the analysis cohort with true-positive result for radiomic analysis (RA) and the convolutional neural network (CNN). Histologic analysis after MRI-guided vacuum-assisted biopsy revealed invasive lobular cancer (G2, pT3 [54 mm], pN3a [13/14], cM0). *A*, First postcontrast subtracted and, *B*, T2-weighted turbo spin-echo images. The green border indicates the region seen by the deep learning algorithm and is magnified in *C–G*. *C*, Precontrast T1-weighted image, *D*, first postcontrast image, and, *E* and *F*, intermediate- and late-phase contrast image. *G*, First postcontrast subtracted image. *H*, At segmentation performed by the radiologist (yellow), a second distant lesion (blue) was analyzed independently; background is red. The segmentation mask images were used for the RA approaches, but not for the CNN, which only got a subset of three rectangular patches (*C–G*) as inputs. Lesion correctly classified as malignant by RA, CNN, and the radiologist.

a momentum of 0.9 and a decaying learning rate starting at 0.001 (decreasing by a factor of 0.05 every seven epochs).

Because the network was pretrained on color images, it expects three input channels. To determine which subset of the seven available sequences should be fed in, all 35 possible three-combinations were tested.

Statistical Analysis

For sample size calculation, we used results from a study (7) regarding RA for classification of lesions found on unenhanced breast MRI images. In that study, a total of 127 lesions were included, yielding an area under the curve (AUC) of 0.85. To detect an improvement of AUC of 0.05 at an α error of .05 and β error of .2, for an equal allocation of patients with benign and malignant lesions, a sample size of at least 702 lesions was deemed necessary (15). However, because we dealt with contrast-enhanced breast MRI and intended to use CNN rather than RA alone, we planned to include at least 1000 enhancing lesions.

To compare the accuracies of the algorithms and the radiologist readings, the respective sensitivity and specificity values were calculated on the basis of a cut-off value that minimizes the metric $m = (1 - \text{sensitivity})^2 + (1 - \text{specificity})^2$ (Figs 2, 3). The AUC was calculated for the respective receiver operating characteristics on the basis of the given numeric value of the algorithm and the respective BI-RADS categories. To do the latter, BI-RADS categories 4–6 were considered test-positive and the remaining were considered test-negative.

To assess the dependency on the size of the underlying dataset, both variants of RA (principal component analysis and L1 regularization) and the CNNs were retrained on a dataset of half the size (224 of 447 patients and 647 of 1294 lesions, of which 393 lesions were malignant), and AUC were re-analyzed. Splitting for cross validation was performed as before.

Standard deviations and confidence intervals were calculated by using bootstrap analysis with 100 000 fold resampling as in Litjens et al (16) (Fig E1 [online]). Significance level for pairwise comparisons between the algorithms was set to .05/6 to adjust for six pairwise comparisons according to Bonferroni (17).

Confidence interval calculation for sensitivity and specificity was based on normal approximation of the binomial distribution.

Because the majority of enhancing lesions larger than 2 cm were malignant, an additional sensitivity analysis was performed for lesions smaller than 2 cm (n = 823).

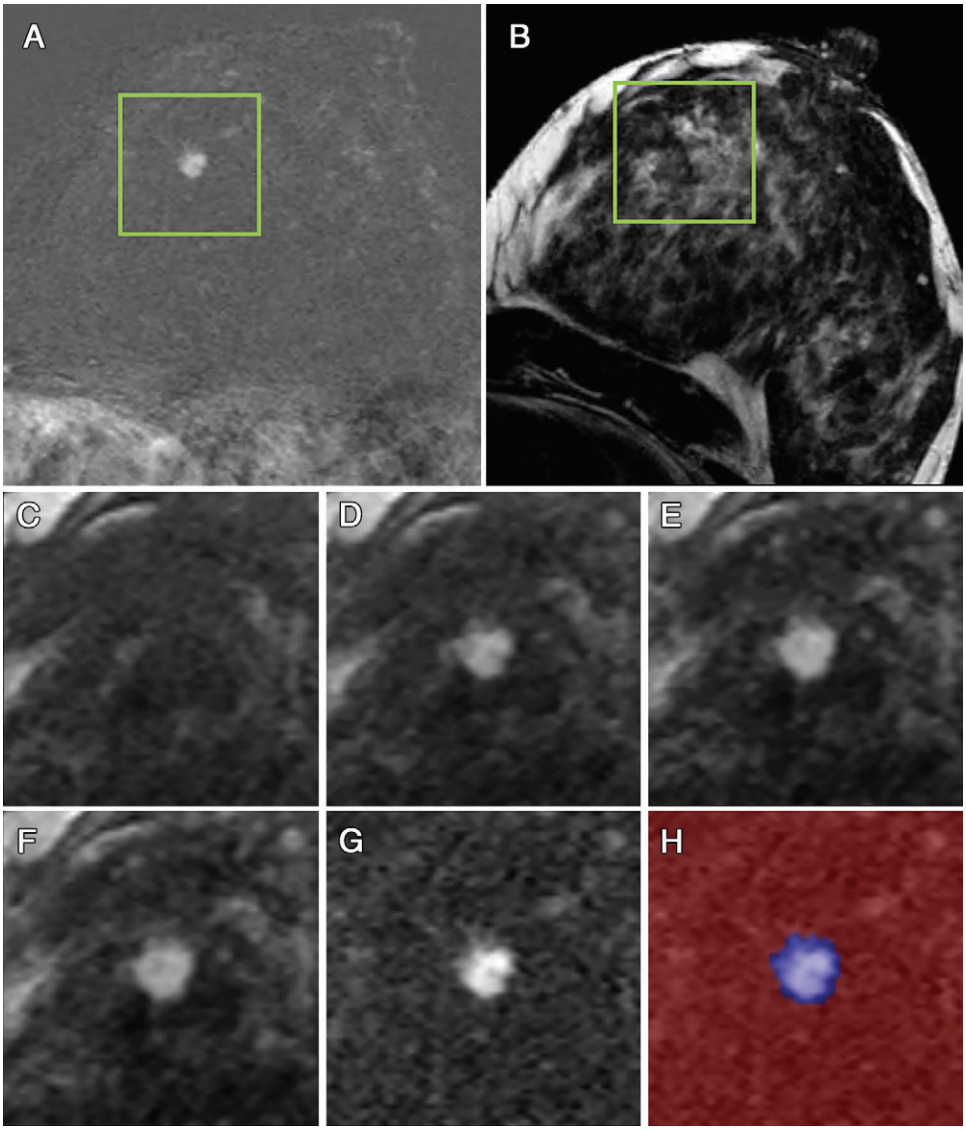

Figure 5: Example of a breast lesion in a 46-year-old female patient in the analysis cohort with a false-negative result by radiomics analysis and the convolutional neural network. Histologic analysis after MRI-guided vacuum-assisted biopsy revealed invasive lobular cancer (G2, pT1b, pN0, cM0). *A*, First postcontrast subtracted and, *B*, T2-weighted fast-spin-echo images. The green border in *A* and *B* indicates the region observed by the deep learning algorithm and is magnified in *C–G*. *C*, Precontrast T1-weighted image; *D*, first postcontrast image; *E* and *F*, intermediate- and late-phase contrast images, respectively; *G*, first postcontrast subtracted image; and, *H*, segmentation performed by the radiologist (blue). The lesion was incorrectly classified by both radiomic analysis and convolutional neural network as benign, but classified correctly by the radiologist as malignant.

Results

The final analysis cohort consisted of bilateral breast MRI data sets from 447 patients (mean age, 66.0 years ± 10.3 [standard deviation]; age range, 26.7–82.0 years) who underwent MRI for indications detailed in Figure 1; patient demographics are given in Table 1.

A total of 1294 enhancing lesions, 507 (39.2%) benign and 787 (60.8%) malignant, were identified and segmented, which yielded an average of 1.4 different types of enhancing lesions per breast. A description of the lesion types is provided in Table 2. Median size of all enhancing lesions was 15 mm ± 20; malignant lesions, 19 mm ± 22; invasive cancers, 16 mm ± 15; and ductal carcinoma in situ, 34 mm ± 29. The median size of benign lesions was 11 mm ± 13 (Fig E2 [online]).

RA by means of L1 regularization led to a smaller subset of 47 relevant image features. Of those, six are texture and statistics features derived from the T2-weigthed image and one is a shape feature (sphericity), whereas the remaining 39 features were texture and statistics features based on the dynamic sequences. Even after penalizing the use of redundant features, the highest interfeature correlations were found for the mean absolute deviation, the robust mean absolute deviation, and the interquartile range of pixel intensities in the subtraction image, more or less denoting tumor enhancement inhomogeneity. Similar blocks were found for degree of enhancement in early and late phase images, providing information on enhancement kinetics. The chosen features for RA and principal component analysis are qualitatively similar (details for both radiomic approaches are in Appendix E1 [online]).

AUC, sensitivity, specificity, and the corresponding 95% confidence intervals are provided in Table 3.

Training of the CNN took 1350 seconds. Prediction of a single lesion took from 66 msec to 528 msec, depending on the number of sections that contained the lesion. Among the seven possible images to feed into the three input channels of the CNN, the sequences that provided the best results were the precontrast and the first and third postcontrast dynamic series.

Breast radiologists' readings yielded an AUC of 0.98 ± 0.01. Example images of correctly and incorrectly classified breast lesions are in Figures 4–6 and Figure E3 (online). None of the three malignant lesions that had been classified as false-negative findings by the radiologist's readings had been correctly identified by either of the three computer algorithms.

No statistically significant differences (P = .04) were observed for the AUCs of two different approaches of RA, RA

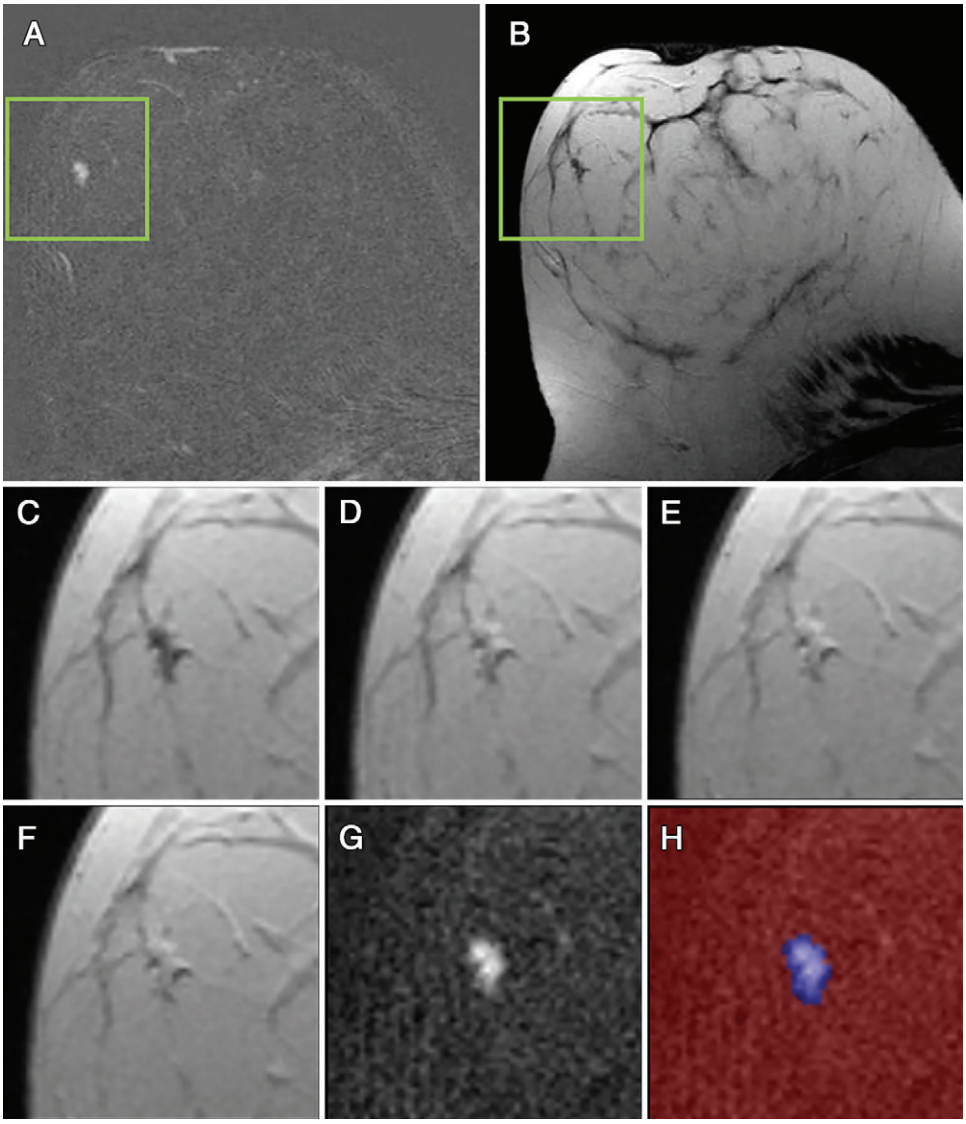

Figure 6: Example of a breast lesion in a 52-year-old female patient in the analysis cohort with a false-negative result by radiomic analysis and a true-positive result by the convolutional neural network. MRI-guided vacuum-assisted biopsy revealed ductal carcinoma in situ. *A*, First postcontrast subtracted and, *B*, T2-weighted fast-spin-echo images. The green border indicates the region observed by the deep learning algorithm and is magnified in figures *C–G*. *C*, Precontrast T1-weighted image; *D*, first postcontrast image; *E* and *F*, intermediate- and late-phase contrast image, respectively; *G*, first postcontrast subtracted image; and, *H*, segmentation as performed by the radiologist (blue). The lesion was incorrectly classified by radiomic analysis as benign, but correctly classified by convolutional neural network and by the radiologist as malignant.

with L1 regularization (AUC, 0.81), and RA with principal component analysis (AUC, 0.78). AUC of CNN (AUC, 0.88) was significantly higher compared with both RA approaches (both comparisons, $P < .001$), but still significantly lower compared with the human readers (AUC, 0.98; $P < .001$) (Table 3; Figs 2a, 3).

Essentially the same results were observed when restricting the analysis to lesions smaller than 2 cm (Fig 2b).

For RA with L1 regularization and RA with principal component analysis, reducing the number of training lesions by half did not lead to significant differences regarding AUC compared with the full training cohort ($P = .05$ for RA and L1 regularization [AUC, 0.80] for training on half- vs full-size cohort; $P = .06$ for RA and principal component analysis [AUC, 0.78]). For CNN, AUC differed significantly when training was compared with half- versus the full-size cohort (AUC, 0.83; $P = .01$).

Discussion

Our analysis of breast MRI studies confirm that a radiomic signature exists that encodes lesion malignancy, which is extractable by RA and deep learning algorithms. However, in spite of the efforts that went into tuning the two radiomic approaches (L1 regularization or principal component analysis), radiomics diagnostic accuracies were lower than that of the CNN, with an AUC of 0.78 and 0.81 for L1 regularization or principal component analysis, respectively, compared with 0.88 for CNN ($P < .001$ for both comparisons). Although the diagnostic accuracies of the RA and the CNN could be considered to be in a clinically acceptable range (18–21), they were both far from matching the performance of the breast radiologists, who yielded an AUC of 0.98. It should also be noted that the diagnostic performance for the machine learning algorithms refers to the classification of lesions that were identified by a radiologist; accordingly, the calculated sensitivity of the algorithms reflects their performance for correctly classifying a preidentified malignant lesion as malignant, but it does not include their performance in finding or detecting a malignant lesion.

Our results suggest that CNNs appear to be the candidate with the more promising development perspective for classification of enhancing lesions: Whereas the performance of both RA approaches did not improve after enlarging the available training data set, with similar AUCs observed after training with the half-size versus the full-size cohort (0.80 vs 0.81 for L1, respectively), AUC of the CNN algorithm did improve significantly from 0.83 to 0.88. This implies that RA exhibit a so-called saturation curve of their attainable accuracies;

possibly, the amount of information encompassed by a fixed set of handmade radiomic features is incapable of differentiating more subtle differences between malignant and benign breast lesions. However, CNNs, with their more complex and easily expandable structure, might be able to mimic the elusive and subconscious process that occurs when a radiologist interprets MR images.

Accordingly, there is reason to assume that inclusion of even more data into our CNN model, and/or more sophisticated data augmentation methods such as generative adversarial neural networks (22), will further improve the results of our CNN model.

Our results are plausible when comparing them with other fields of medical imaging, such as interpretation of skin lesions, in which CNNs have achieved diagnostic accuracy levels that have been unattainable by using RA (23). In the field of radiology, Kooi et al (5) demonstrated that CNNs, trained on a large data sets of screening mammograms, are superior to state-of-the-art computer-aided detection software (ie, algorithms that use principles of RA). In our study, a relatively low number of cases was sufficient to achieve a relatively high diagnostic accuracy of the CNNs. This is likely because we used high-quality annotated data for training, provided by an experienced breast radiologist.

The diagnostic accuracy achieved by breast radiologists in our cohort reflects the upper limit of the range of diagnostic accuracies reported for breast MRI, which is partly because we worked on an artificial cohort with a high prevalence of malignant lesions; it is well established that diagnostic accuracy levels will vary (ie, be higher) with a high prevalence of breast cancer. Another reason for the relatively high diagnostic accuracy of radiologists in our study is because our department receives high volumes of referrals for breast MRI, and so these readers obtain ample experience in interpreting breast MR images.

Our study had several limitations. First, sensitivity in our study refers to the correct classification of lesions that had been preidentified and segmented by a radiologist. Second, our MR examinations were acquired with a standardized protocol, whereas across different institutions pulse sequence protocols vary; accordingly, the trained algorithms may not achieve the same accuracy on data from other acquisition protocols. Finally, advanced data augmentation by using generative adversarial networks (22) could help to adapt the network structure to a more suitable three-dimensional approach to incorporate all of the available imaging sequences, and our algorithm only used a subset of three as described.

To conclude, a CNN was superior to radiomics algorithms for classification of enhancing lesions at multiparametric breast MRI. Even with a limited training data set, the diagnostic accuracy achieved with CNN seems to reach clinically acceptable levels. Whereas the CNN was inferior to breast radiologists, the CNN approach may have the potential to improve its performance with the future availability of additional and larger datasets.

Author contributions: Guarantors of integrity of entire study, S.S., C.K.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, all authors; clinical studies, D.T., H.S., C.K.; experimental studies, D.T., C.H., H.S., D.M.; statistical analysis, D.T., S.S., C.H., H.S.; and manuscript editing, all authors

Disclosures of Conflicts of Interest: **D.T.** disclosed no relevant relationships. **S.S.** disclosed no relevant relationships. **C.H.** disclosed no relevant relationships. **H.S.** disclosed no relevant relationships. **D.M.** disclosed no relevant relationships. **C.K.** disclosed no relevant relationships.

References

1. Sardanelli F, Boetes C, Borisch B, et al. Magnetic resonance imaging of the breast: recommendations from the EUSOMA working group. Eur J Cancer 2010;46(8):1296–1316.
2. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology 2016;278(2):563–577.
3. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5(1):4006.
4. Li H, Zhu Y, Burnside ES, et al. MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, Oncotype DX, and PAM50 gene assays. Radiology 2016;281(2):382–391.
5. Kooi T, Litjens G, van Ginneken B, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal 2017;35:303–312.
6. Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer. Invest Radiol 2017;52(7):434–440.
7. Bickelhaupt S, Paech D, Kickingereder P, et al. Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography. J Magn Reson Imaging 2017;46(2):604–616.
8. Kuhl CK, Strobel K, Bieling H, Leutner C, Schild HH, Schrading S. Supplemental breast MR imaging screening of women with average risk of breast cancer. Radiology 2017;283(2):361–370.
9. Oliphant TE. Python for scientific computing. Comput Sci Eng 2007;9(3):10–20.
10. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 2010;29(6):1310–1320.
11. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res 2017;77(21):e104–e107.
12. Haarburger C, Langenberg P, Truhn D, et al. Transfer learning for breast cancer malignancy classification based on dynamic contrast-enhanced MR images. In: Maier A, Deserno T, Handels H, Maier-Hein K, Palm C, Tolxdorff T, eds. Bildverarbeitung für die Medizin 2018. Berlin, Germany: Springer, 2018; 216–221.
13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, June 27–30, 2016. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2016; 770–778.
14. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–252.
15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.
16. Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6(1):26286.
17. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. Stat Med 2014;33(11):1946–1978.
18. Leach MO, Boggis CR, Dixon AK, et al. Screening with magnetic resonance imaging and mammography of a UK population at high familial risk of breast cancer: a prospective multicentre cohort study (MARIBS). Lancet 2005;365(9473):1769–1778 [Published correction appears in Lancet 2005;365(9474):1848.] https://doi.org/10.1016/S0140-6736(05)66481-1.
19. Kriege M, Brekelmans CT, Boetes C, et al. Efficacy of MRI and mammography for breast-cancer screening in women with a familial or genetic predisposition. N Engl J Med 2004;351(5):427–437.
20. Warner E, Plewes DB, Shumak RS, et al. Comparison of breast magnetic resonance imaging, mammography, and ultrasound for surveillance of women at high risk for hereditary breast cancer. J Clin Oncol 2001;19(15):3524–3531.
21. Kuhl C, Weigel S, Schrading S, et al. Prospective multicenter cohort study to refine management recommendations for women at elevated familial risk of breast cancer: the EVA trial. J Clin Oncol 2010;28(9):1450–1457.
22. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, Volume 2, Montreal, Canada, December 8–13, 2014. Cambridge, Mass: MIT Press, 2014; 2672–2680.
23. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–118.